

DOCUMENT RESUME

ED 418 862

SE 061 341

AUTHOR Boone, William J.
TITLE Systemic Reform: Seeking To Understand the Consequences..
Test Item Measurement in Bridging Selected Steps Which Help
in the Measurement and Monitoring of Systemic Reform.
PUB DATE 1998-04-00
NOTE 13p.; Paper presented at the Annual Meeting of the National
Association for Research in Science Teaching (71st, San
Diego, CA, April 19-22, 1998).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Standards; *Educational Change; Elementary
Secondary Education; *Program Evaluation; Science Education;
*Science Teachers; *Science Tests; Scientific Literacy;
Standards; *Student Evaluation
IDENTIFIERS National Assessment of Educational Progress; *Systemic
Educational Reform

ABSTRACT

An important component of the Bridging study involves the collection of achievement data using sets of derived test items from the National Assessment of Educational Progress (NAEP), as well as new items which reflect state and national standards in science. From these test data, student achievement measures are computed, then these measures are in turn used for subsequent analyses which utilize a varied array of data types such as attitude, race, gender, and socioeconomic status. A number of important issues involving test items are discussed including the monitoring of test items, the targeting of test items, and the use of statistical methods which do not penalize students and allow flexibility with tests. Details are included on the development of the test bank and a delineation of the steps taken to optimize the development and use of the test item bank. (DDR)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

Systemic Reform: Seeking to Understand the Consequences

Test Item Measurement in Bridging Selected Steps Which Help in the Measurement and Monitoring of Systemic Reform

William J. Boone
School of Education 3068
Indiana University
Bloomington, IN 47401

1-812-856-8132
wboone@indiana.edu

1998 NARST Conference

PERMISSION TO REPRODUCE AND
USE THIS MATERIAL HAS
BEEN GRANTED BY

W. Boone

THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to
improve reproduction quality.

• Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

This material is based upon research supported by the National Science Foundation. Any
opinions, findings, and conclusions or recommendations in this publication are those of the
authors and do not necessarily reflect the views of the National Science Foundation.

BEST COPY AVAILABLE

Test Item Measurement in Bridging- Key Steps to Accurately Calculating Achievement Measures

Introduction:

An important component of the Bridging study involves the collection of achievement data using sets of NAEP derived test items, as well as new items which reflect state and national standards in science. From these test data, student achievement measures are computed, and these measures are then in turn used for subsequent analyses which utilize a varied array of data types (e.g., attitude, race, gender, social economic status). In this presentation, a number of important issues involving test items will be discussed- 1) the monitoring of test items, 2) the targeting of test items, and 3) the use of statistical methods which do not penalize students and allow flexibility with tests. All of these issues are critical for the evaluation of students over a period of time.

The Issue:

A variety of studies investigating equity issues have utilized achievement data collected over a period of time. This is often done to monitor the expansion and closing of achievement gaps among groups of students. For instance, the gap in achievement between African American males and white males, or the gap in achievement of Hispanics and African Americans. The techniques used to document and understand these gaps have been outlined in a range of studies and presented in a variety of published papers. Analysis steps within science education seem to follow a set pattern. First- an argument is presented which outlines a rationale for the set of items presented in achievement tests to subjects. This argument is commonly built around the use of expert judges, the calculation of reliability values, and some discussion of validity. The second common step involves the use of raw scores to calculate the performance of student abilities. The third step is the use of a statical procedures to compare students throughout the course of an investigation. Commonly the same test is used as a function of time.

The testing issues considered in Bridging present an improved technique of test data collection. For instance, instruments need not be viewed as static once a project begins, but rather instruments must be flexible (instruments can be improved over the course of a project). This and other issues are of particular importance when one considers that in studies such as Bridging, new information and guidance results from each year of data collection.

Selection, Development and Targeting of Test Items:

In many equity studies, achievement tests need to be developed and administered to students. Often those test item banks are developed through the use of expert judges, teachers, administrators, and national data. This technique, used in Bridging (and used in other studies) is a realistic first step. However, many projects overlook two critical issues when developing and using an item bank. First- items are often not carefully selected to measure the traits which are to be monitored. Secondly, following the first round of data collection, many projects do not review the test as a whole, and specific test items are not reviewed in terms of how well students' achievement is being gauged. Third-projects often do not attempt to improve the test through the development of new items, or the selection of added test questions using existing well piloted

items.

In the Bridging study a number of steps were taken to optimize the development and use of the test item bank.

Step I:

First- the initial Bridging item bank was developed using NAEP items. Only items that were classified as being of "above" the knowledge (or factual level) were utilized. The reason for excluding "knowledge" level questions was simply related to the goal of the state SSI and the goal of measurement with the test items - namely the achievement test was not one which sought an assessment of factual science knowledge, but rather was to be one which would evaluate inquiry. The advantage of using NAEP items is that the items had been piloted, and nationwide "p" values for each item exist. Although "p" values may not always be generalizable to a state, the values do provide some broad guidance with respect to the ease or difficulty of items. It does make sense to utilize experts for the development of test items, however, it seems prudent to make use of well piloted items. Although techniques of spotting and investigating items which may be biased racially and/or biased with respect to gender have been improved in recent years, bias detection techniques used to improve the initial NAEP item drafts, and develop final NAEP items certainly serve as a good starting point for item selection.

Step II:

A critical issue in any evaluation effort is to measure at a variety of time points. However, it is simply not enough to design a measurement device and collect data (without fail) using the identical instrument from time point to time point. It is of great importance to evaluate how well test items selected for the calculation of mean measures are working with the sample of students. In our analysis of the initial 28 item science test developed for Ohio's SSI, and used as a base for this study, we discovered that "difficult" and "easy" items needed to be added to the test. There were a large number of poorly performing students, but not many items targeted at their ability level. For instance, a large number of poorly performing students typically correctly answered a very small number of items (e.g. 0-4 of 28). This meant, for instance, among this group of students, it was difficult to differentiate performance. Likewise many of the best performing students correctly answered the same number of items correctly (e.g. 25-28 of 28). This meant it was difficult to differentiate performance within this group as well. As a result of this issue, items which were quite difficult and quite easy for year 2 testing were added to the science test. When evaluating groups of students, which traditionally have been poorly performing, this step in terms of test item bank maintenance is particularly important- for one must work toward the most accurate gauging of student ability using a reasonable and well targeted set of test items.

Step III:

Removing items from tests which appear to add little to the measurement of students is another key step. To understand this issue, one must consider what an optimal test will look like in

terms of item difficulty. Each item of a test should be uniquely difficult for the tested students. A 30 item test should involve 30 items, each of which has a specific difficulty level. Items should not be at approximately the same level of difficulty. When that is the case, the two items on the test really operate as one, and do not add to what is revealed when evaluating the test data. This step, one in which items are removed if they do not add to the precision of student measures, is of particular importance when one wishes to collect authentic and useful data from students who may not be motivated with respect to test taking, and may not be test savvy. By shortening a test as much as possible through the removal of items which do not differentiate students, a more supportive testing environment is created, and data provides student ability measures of greater precision.

Step IV:

Monitoring Test Items

Monitoring test items in terms of difficulty and ease for students, and the distribution of those items in terms of overlapping measurement, can add greatly to the precision of student measurement. In addition to using the above mentioned techniques, differential item functioning can be used to investigate the operation and interplay of test items. This technique, called DIF (for differential item functioning) can be used to monitor items in terms of students as a function of race and gender. In fact DIF can be used to evaluate the functioning of items in terms of any subgroup. In terms of the Bridging study DIF has been used to specifically evaluate and monitor items as a function of race and gender.

What is DIF and what does it tell you? To understand the issue of differential item functioning, one must first consider what is assumed when a student's test score is used to compare groups of test takers. On a 20 item test, one commonly assumes that the 20 items sample the tested latent trait (e.g. inquiry science) in the same manner (more or less) for all the tested students. This does not mean that all items are of the same difficulty for all students, but what it does mean is that the way in which the items define inquiry science for a student (e.g. Billy) is the same as the way in which the items define inquiry science for any other student (e.g. Mary Jo). The ordering of items from most difficult to most easy is the same for both students. Another way to explain this, is to consider the markings on a yardstick. Although the measure of Billy's total height may be different from the measure of Mary Jo's total height with the yardstick, the ordering of the inch marks on the yardstick is the same no matter who is being measured.

What does this really have to do with achievement measures of students in the Bridging study? In Bridging we are particularly interested the performance of students as a function of gender and race. If items within a test item bank do not function in the same manner among subgroups, then this means that those test items define the latent-trait (the tested variable) in a different manner as a function of subgroups. This can then influence the calculation of students' ability measures, and mean group ability measures.

In order to improve the Bridging item bank, so that potential bias could be monitored and evaluated, a DIF analysis was conducted using the 1995, 1996 and 1997 test data. Plots which

summarize this analysis are attached. The key aspect of these plots are to consider those items which fall outside the 99% confidence interval bands. Those items are ones which may be answered in a slightly different manner by boys and girls or African-Americans and whites. However, if an item does differ as a function of gender one particular year--it does not mean that particular item was easier or harder for boys than girls--rather it simply means the way in which that item defined the trait of science measured by the set of test items may have differed as a function of gender.

DIF Analysis (Comparison of the 1995, 1996 and 1997 science tests)

Items that fall outside the 99% confidence interval bands are items that may define a slightly different scale for the two compared groups. Figures 1-3 present the DIF analysis as a function of gender, while figures 4-6 present the analysis as a function of race.

In general, item 4 of the science test in 1995 and 1996 seemed to measure males and females in a different manner. But in 1997 this difference was not observed. This item involved the issue of inertia.

Item 26 of the 1995 science test (numbered 17 in 1996 and 1997) was built around graph reading and the issue of dinosaurs. Of all the science items evaluated in terms of gender and race, this item exhibited the most consistent DIF, for it was present all 3 years.

Implications for the Observed DIF in Terms of Equity and Racial Issues

Differential item functioning is but one tool by which researchers can carefully monitor the functioning of a test item bank and insure that a measurement device is being used which does not favor one group or another. There can be, as one might expect, differences from year to year in the way in which items define a latent-trait. Due to this issue, it seems most prudent to monitor items before they are removed as the result of DIF analysis. In the case of this study, the dinosaur science item was removed from the calculations of student performance, since DIF analysis revealed questionable functioning of this item for three years.

Step V:

Linear Metrics, Missing Data

Commonly, evaluation efforts use raw scores of student performance on tests to measure students. However, the difficulty with the use of raw scores is that often a non-linear metric is provided. This in turn can greatly effect subsequent calculations. Another problem with the use of raw scores, is that there can be a problem with the issue of missing data. Raw score calculations often force the counting of "items not attempted" as wrong. However, if the goal of a study is to not penalize less savvy test takers (often minority students), then it is critical to use analysis techniques that do not penalize students when a test is not completed. By this we mean, not to count missing data against students. In this study we have utilized IRT techniques (e.g. TIMSS, CTB/McGraw Hill Terra Nova) to calculate student measures on a linear scale.

This has enabled us to utilize parametric equations, and to not penalize students who did not complete the test.

Step VI:

Finally, in terms of carefully measuring students it is important to be able to monitor tests, alter tests, and link tests together. By altering tests, students can be measured with greater precision, and by linking tests one is able to monitor students' growth on one single metric. By utilizing common items across each year (1995 to 1996, 1996 to 1997) students' performance can be expressed on one metric. This is the same technique which is now used by CTB/McGraw Hill to equate differing forms of their newly developed Terra Nova exam.

Conclusion:

Achievement measure are, and should be, an important component of equity studies. However, it is critical to consider some of the steps followed in the Bridging study. By considering the targeting of test items, revising items, monitoring differential item functioning, and linking test item banks- careful measures of students can be made and then utilized in subsequent calculations.

c:\u\discover\narst.98

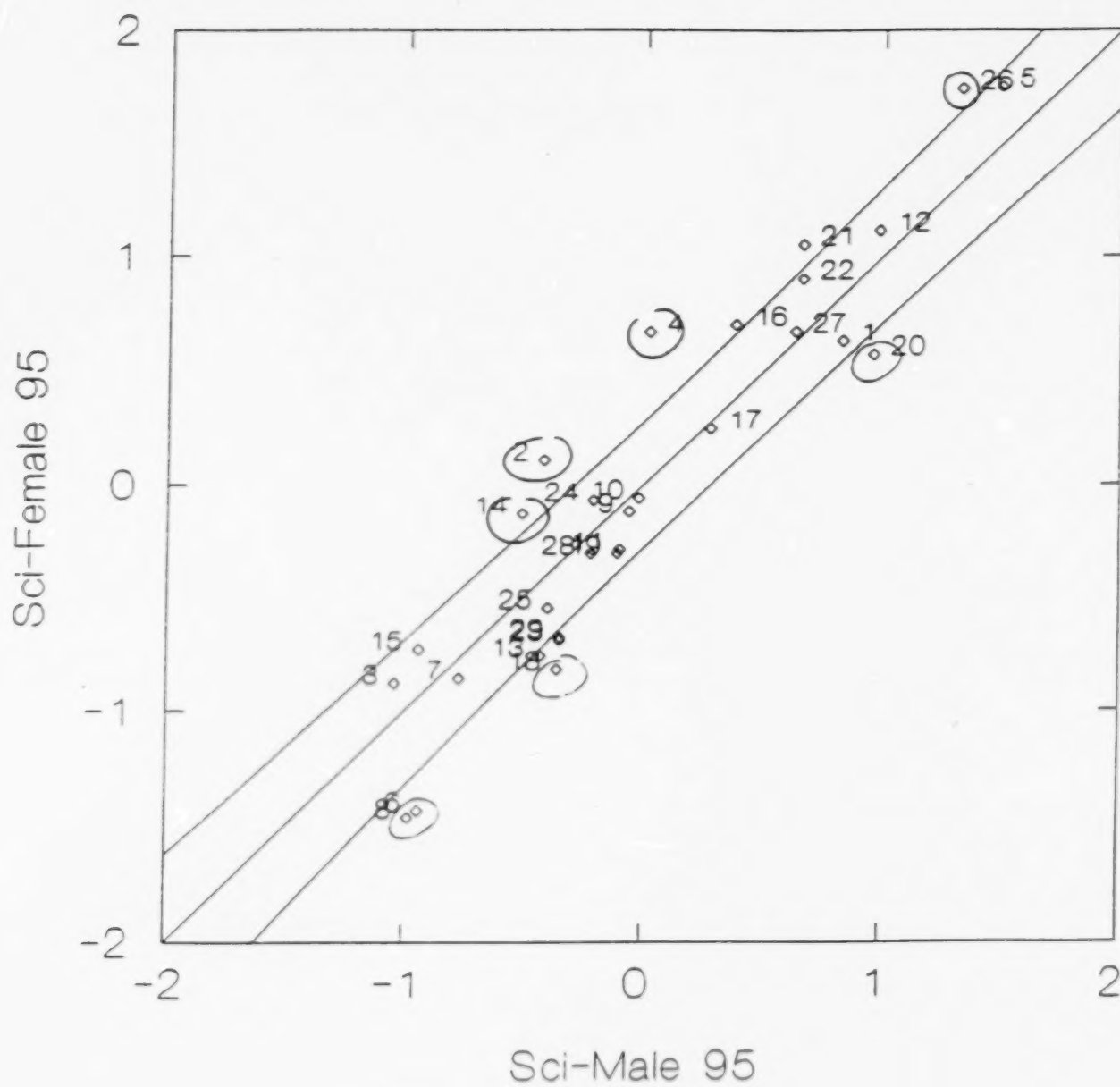


Figure 1

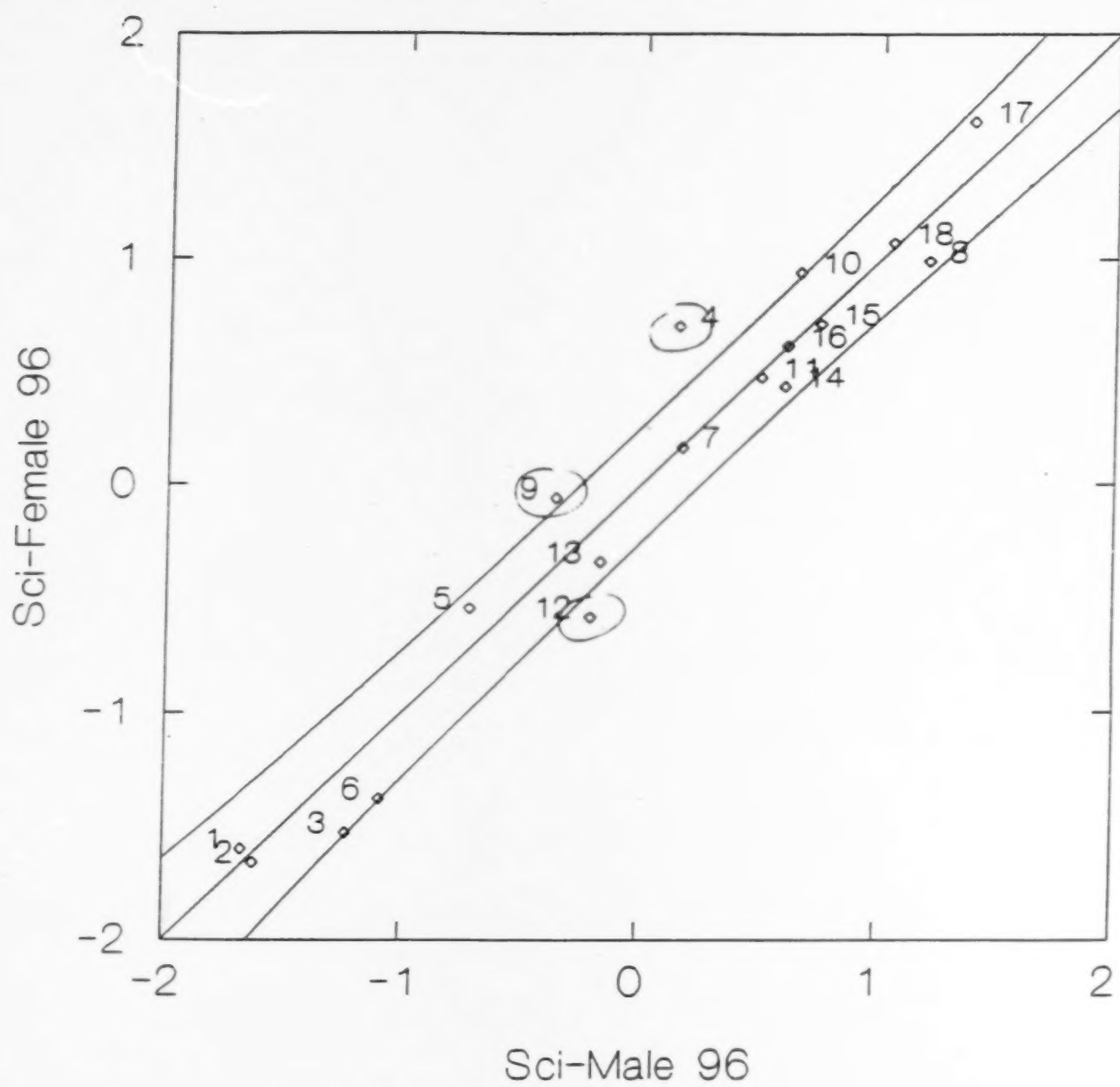


Figure 2

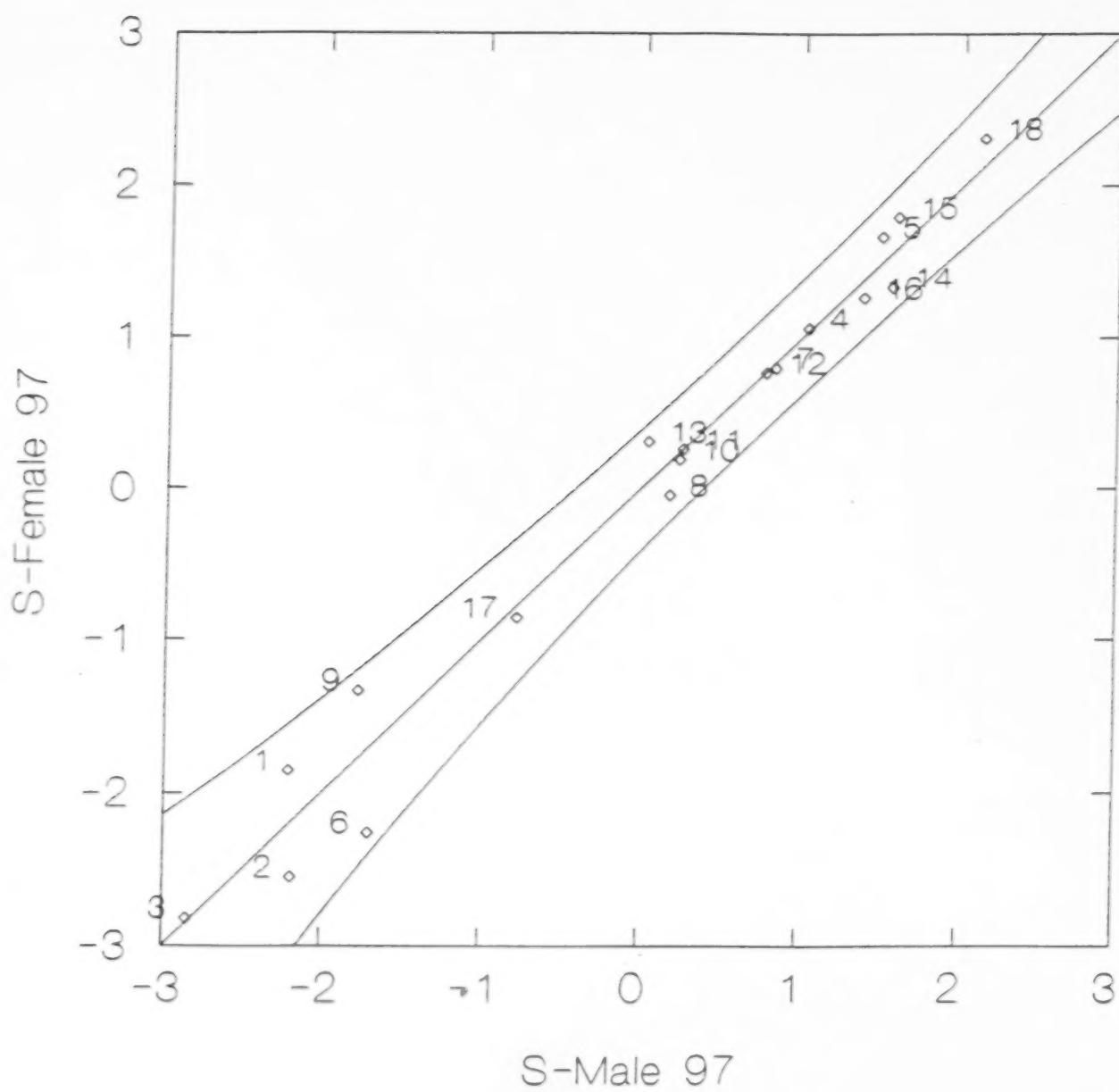


Figure 3

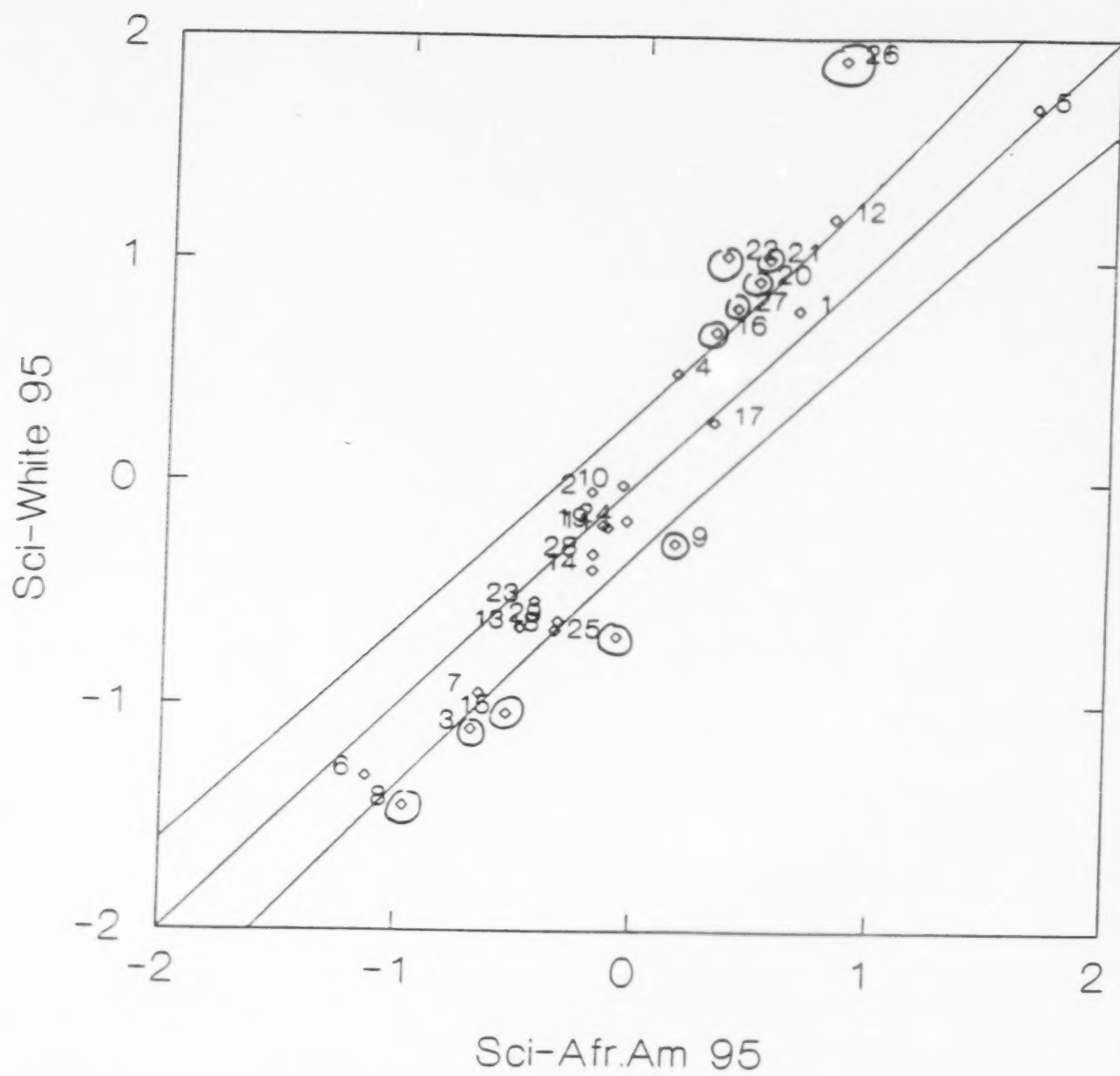


Figure 4

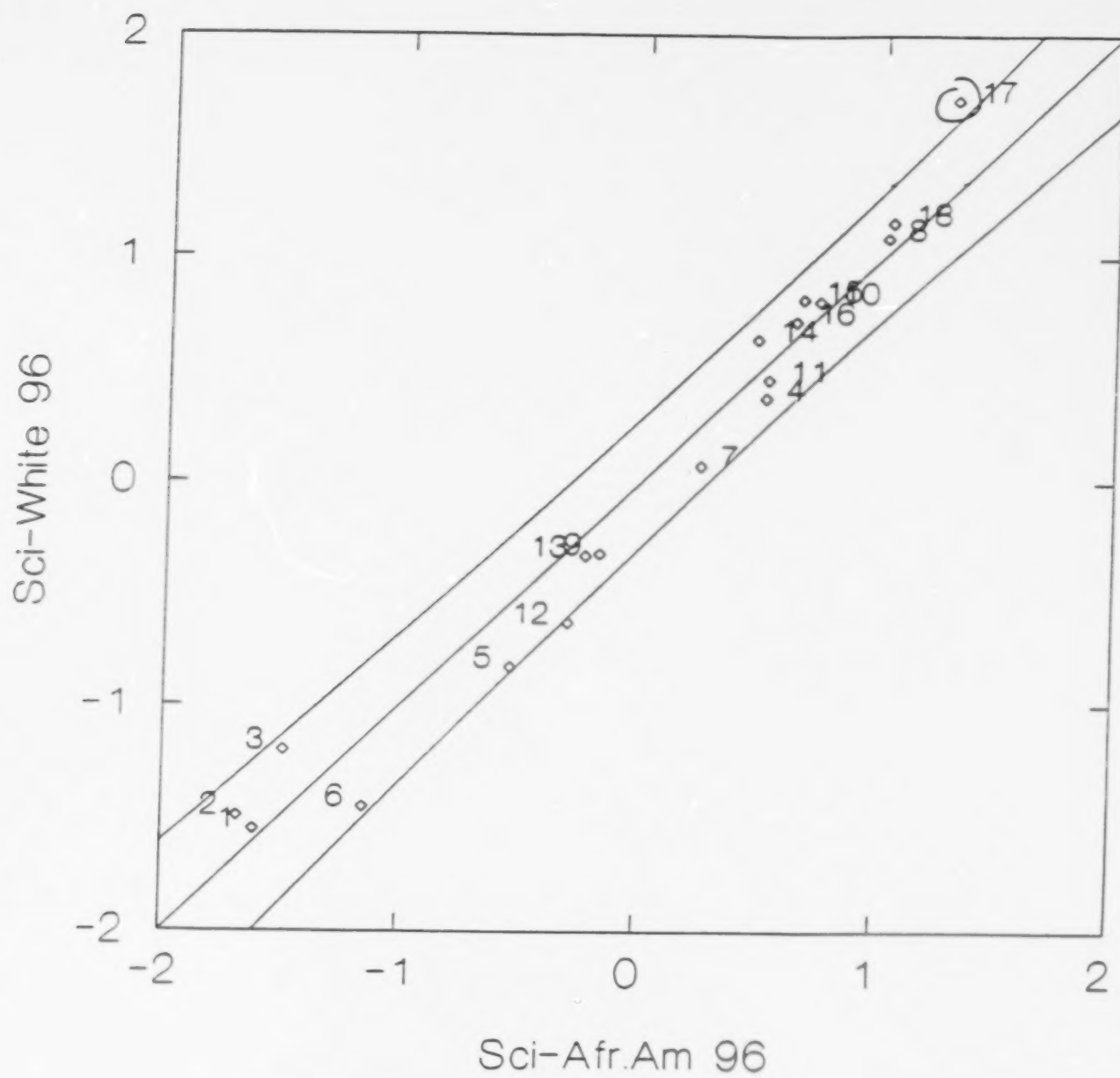


Figure 5

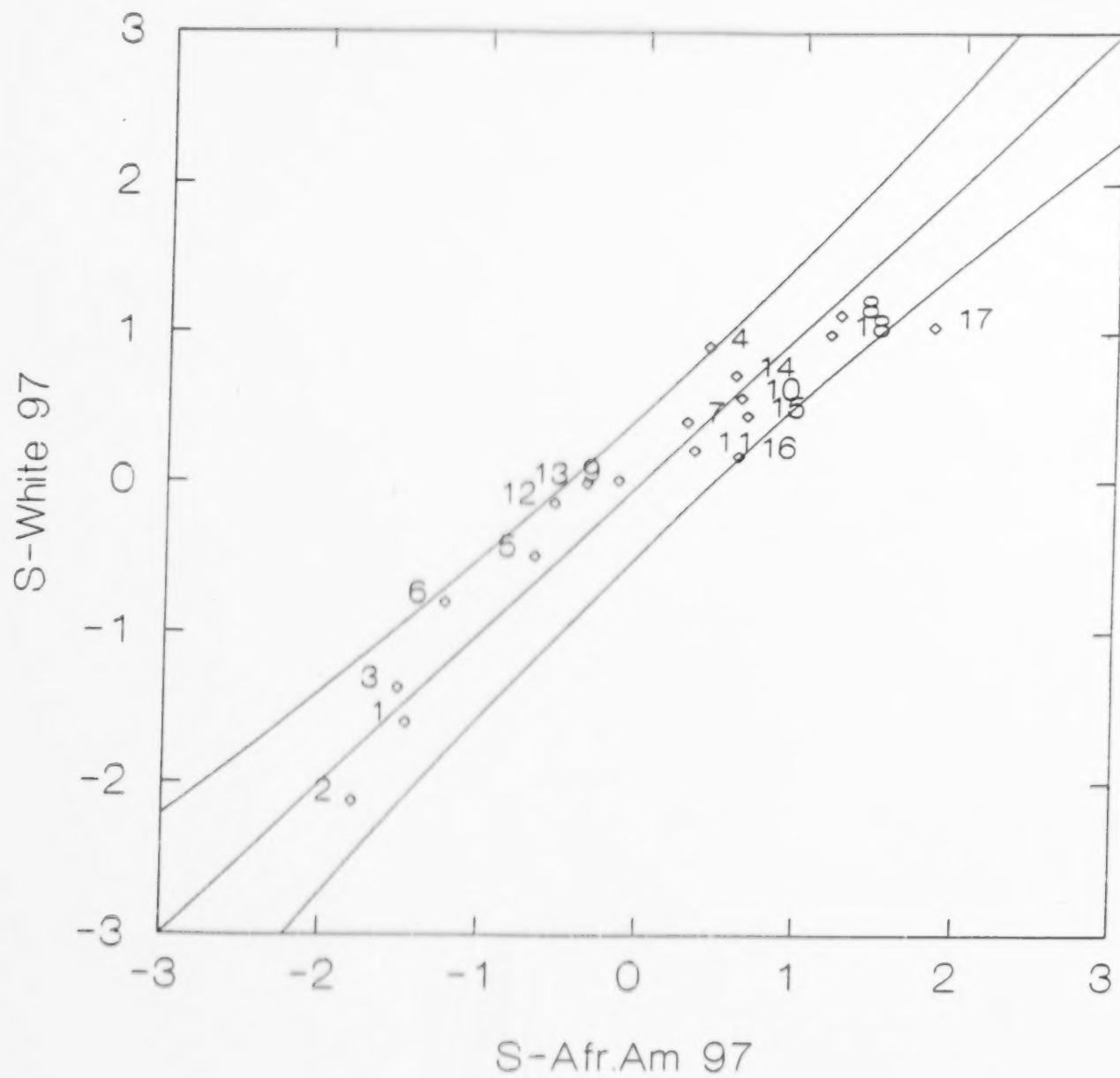


Figure 6